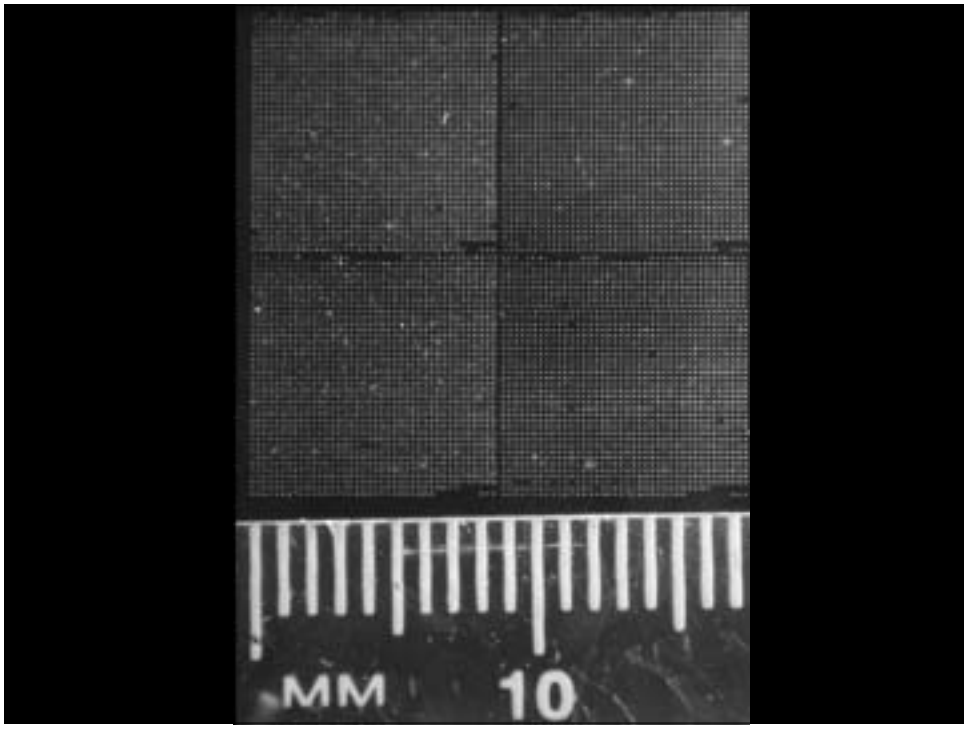
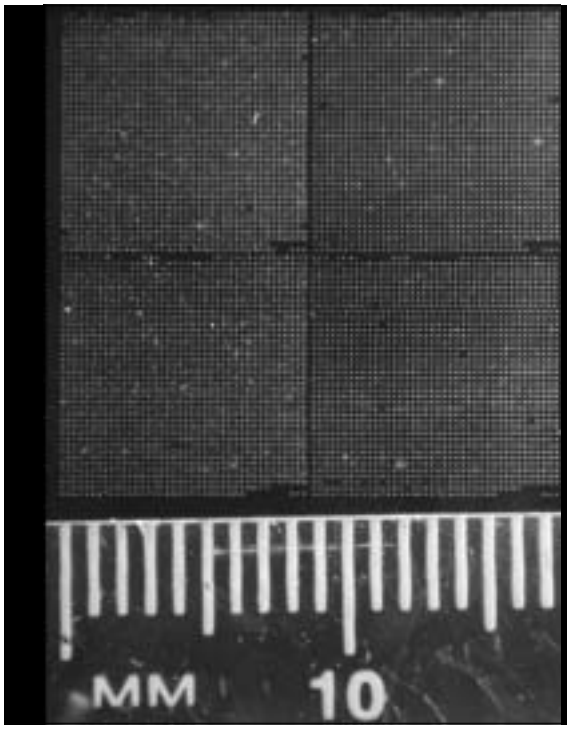
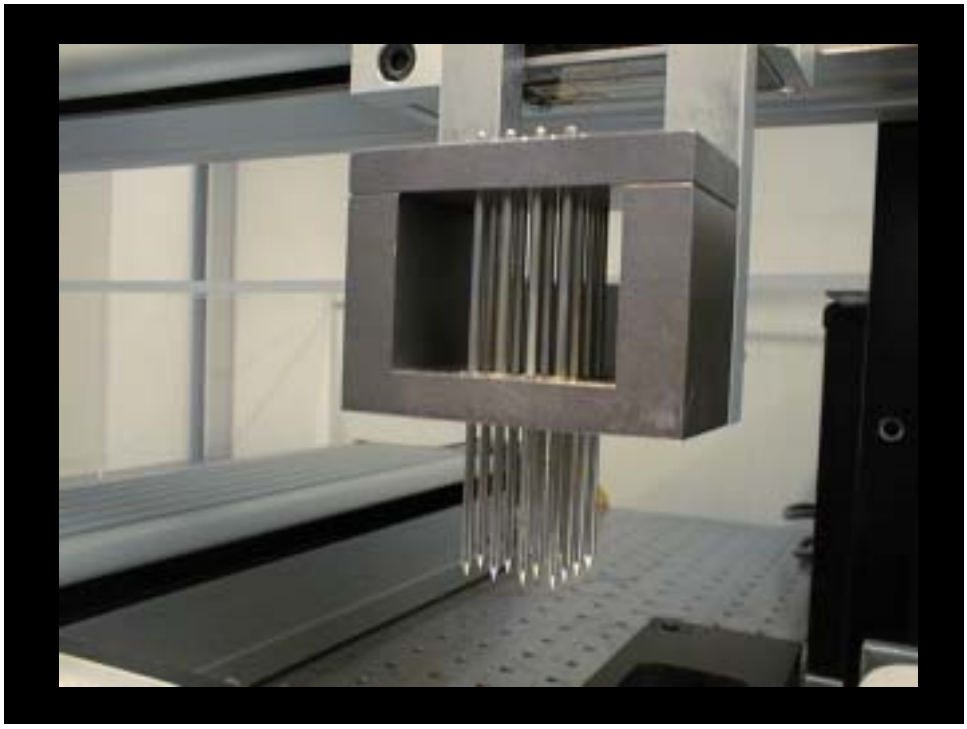
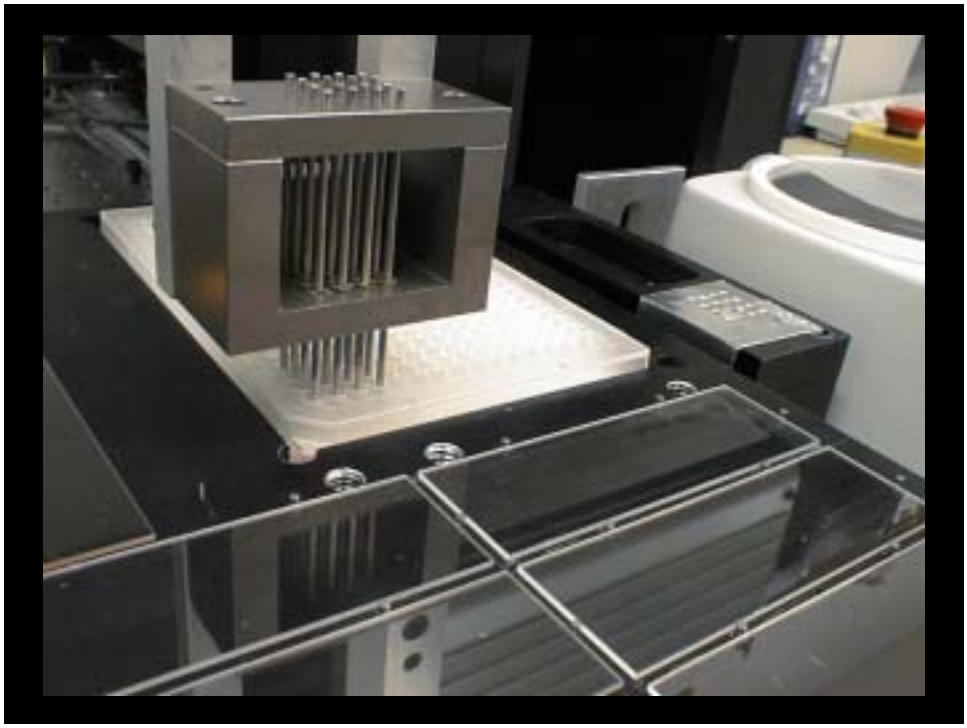


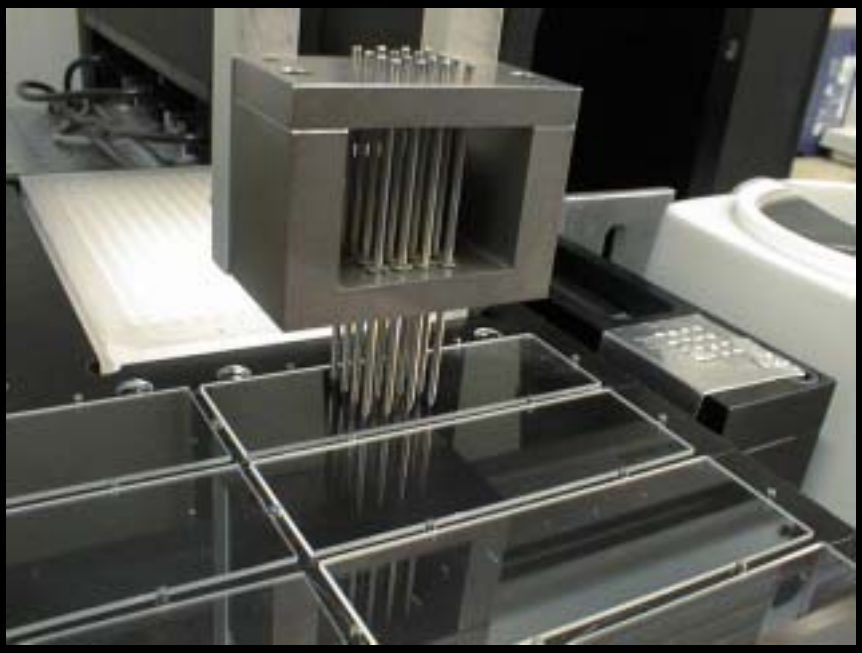


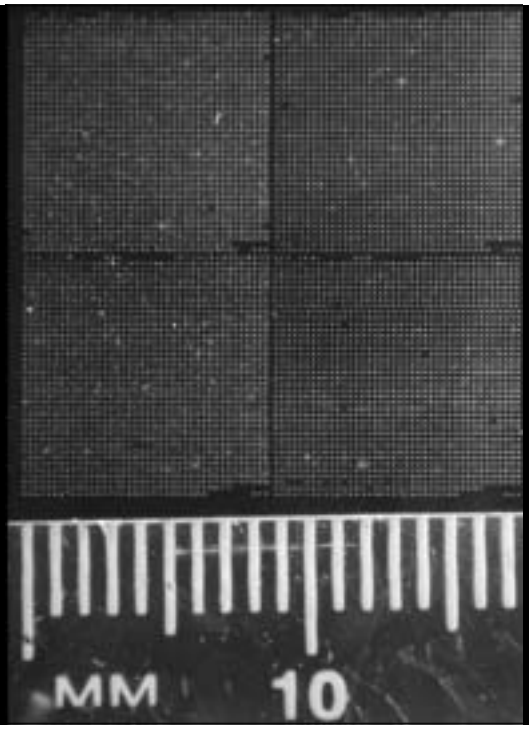
Spotted DNA Arrays

Pre-synthesized DNA (made by PCR or oligo synthesis) is robotically deposited on specially treated glass microscope slides in a grid.









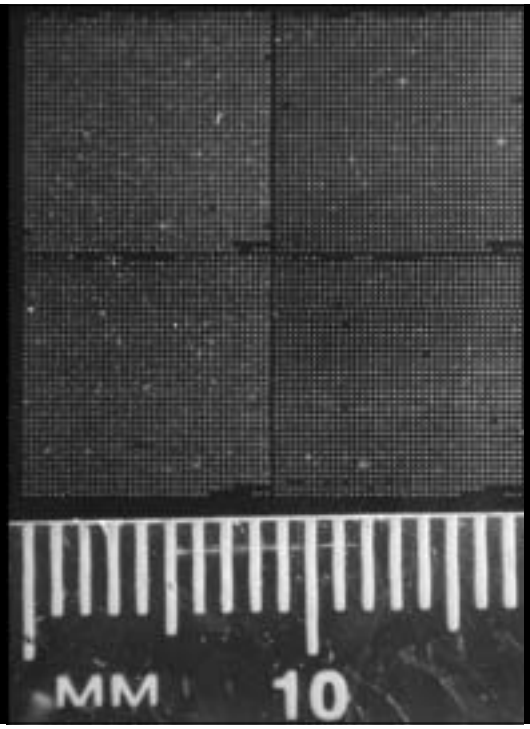
Genome-wide Template for Hybridization Assays

Examples of spotted material:

PCR amplified open reading frames from yeast

Human cDNAs

Synthetic 70-mers from coding sequences of Drosophila



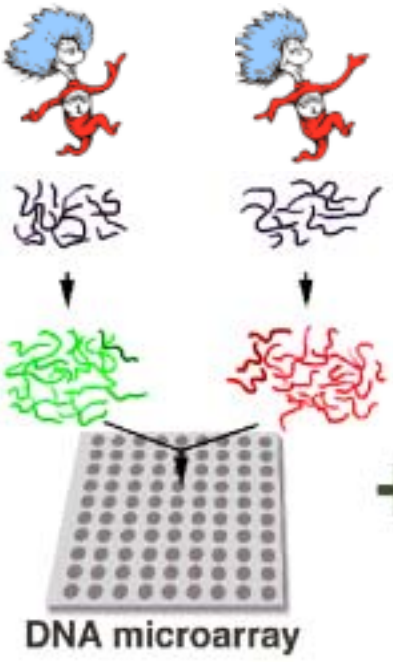
Spotted DNA Arrays

Spots can now be printed with center to center spacing of less than 100um. This permits more than 150,000 spots to be printed on a standard glass slide.

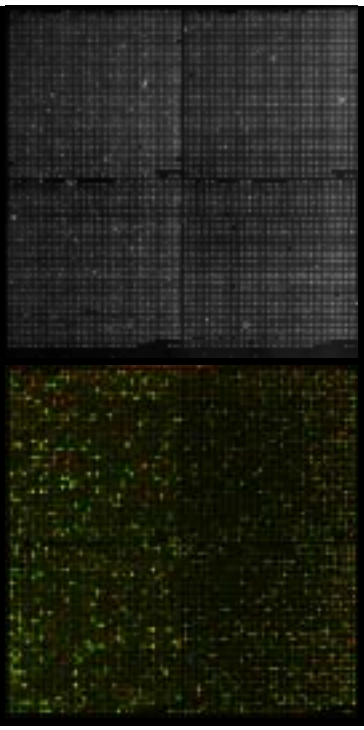
A good robot can now print 50,000 spots on 200 slides in 24 hours

Uses of Arrays

In the most general sense DNA arrays, and other related techniques, make it possible to quantify the composition of complex nucleic acid samples in parallel at high speed and with reasonably high accuracy and (in principle) low cost



Two-color comparative hybridization to arrays measures relative (not absolute) abundance of each



Spotted DNA Arrays

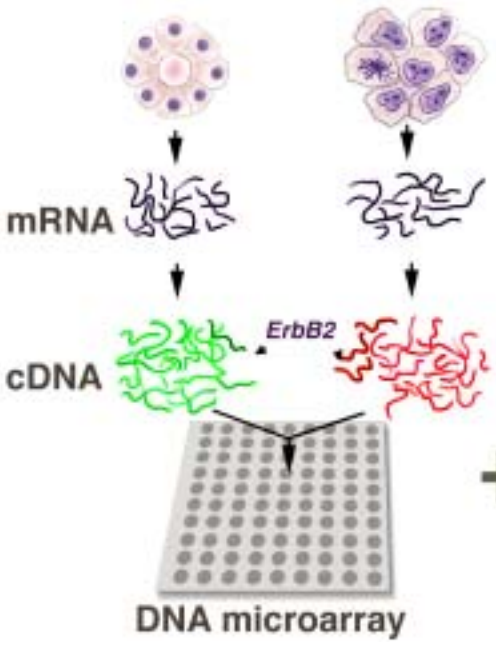
Comparative two-color hybridization to DNA arrays is highly reproducible and sensitive (accurately measure 1.5x changes in species that exist at 1 part in 6 million in the sample)

Uses of Arrays

The biological meaning of the composition of a complex nucleic acid sample depends upon what the sample is and how it was prepared

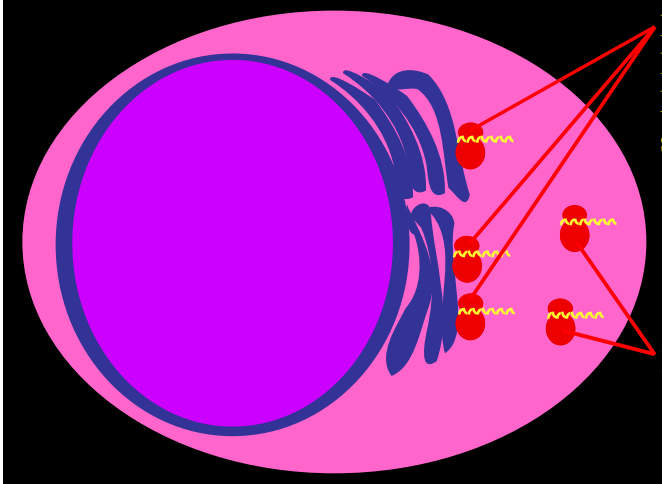
Properties of Biological Systems and Biomolecules That Can Be Studied By Hybridization

- Transcript Abundance
- Karyotype/DNA copy number
- Identity by descent/Genetic Mapping
- Translation, Transcription, Message Decay Rates
- Sub-cellular localization of transcript/gene product
- *In vivo* binding distribution and *in vitro* binding affinities of DNA binding proteins



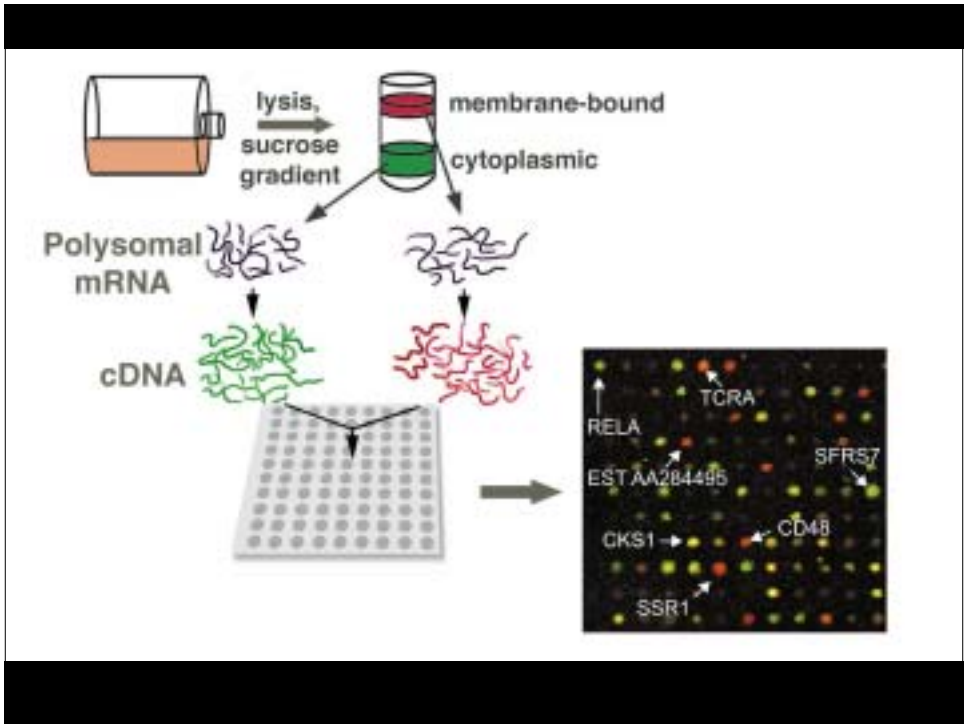
When the hybridized material is cDNA prepared from mRNA isolated from two different biological samples, the result of the hybridization is the measurement of relative transcript abundance between the two samples.

Use of cDNA Microarrays to Identify Genes Encoding Membrane and Secreted Proteins

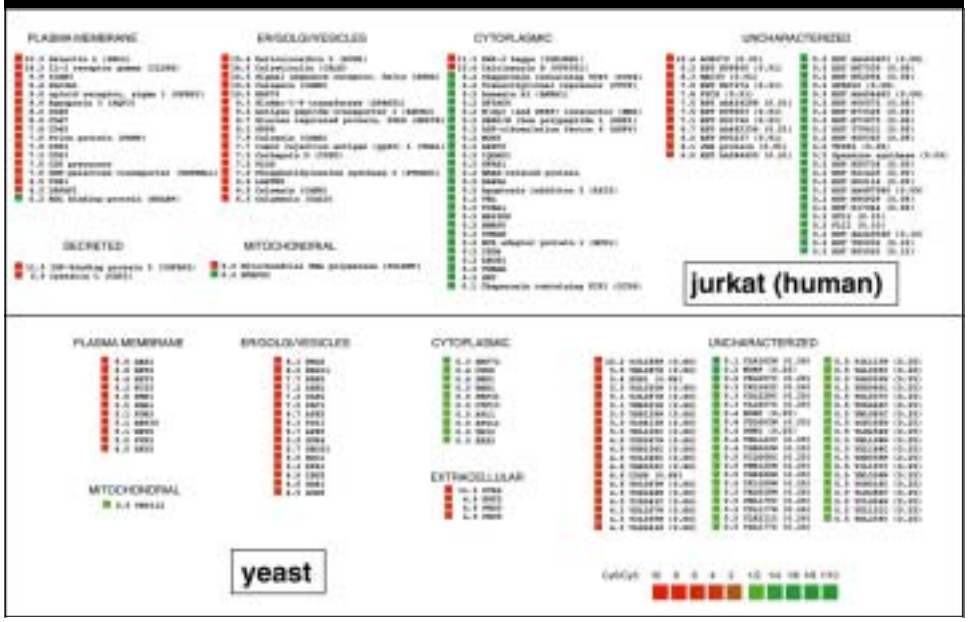


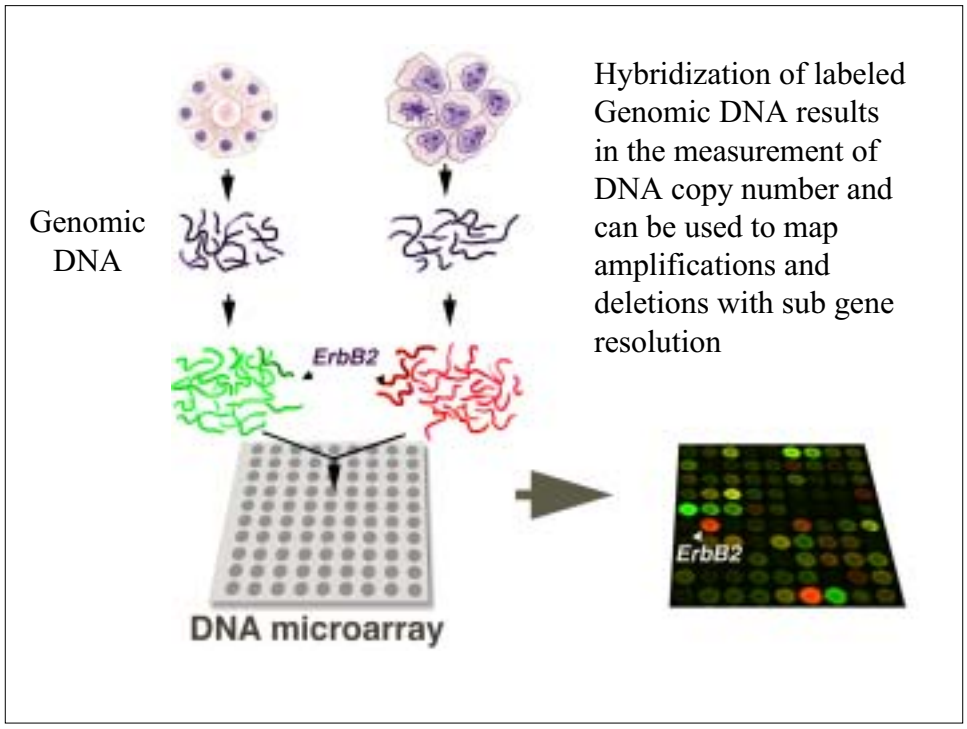
Membrane-bound Polysomes => Membrane and secreted proteins

Free polysomes => Cytoplasmic and nuclear proteins

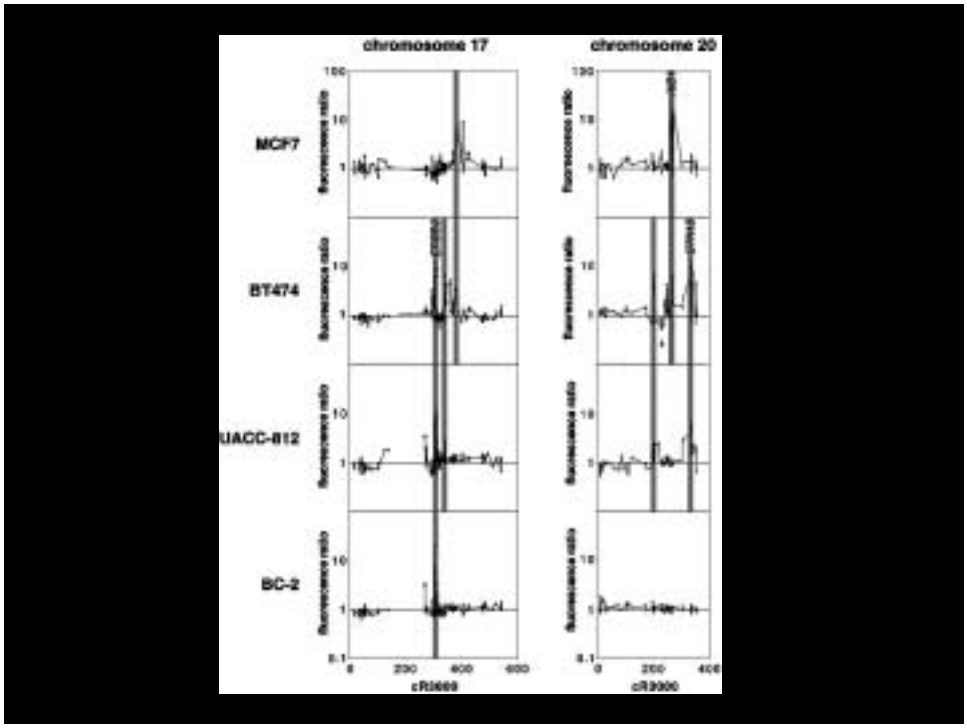


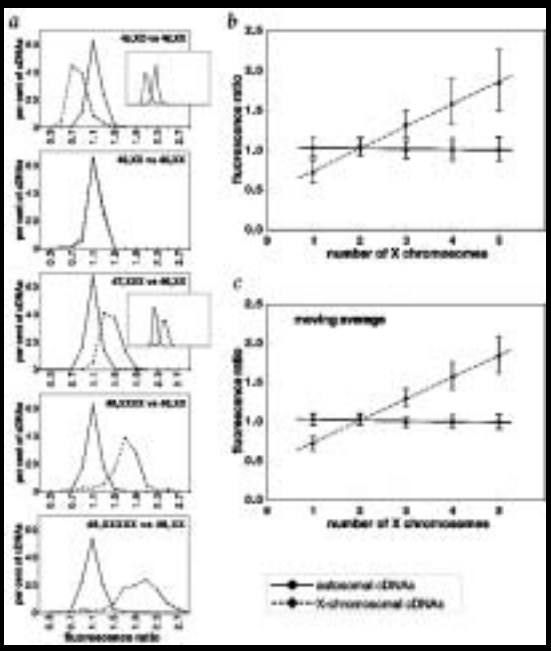
Results of MBP Array Analyses

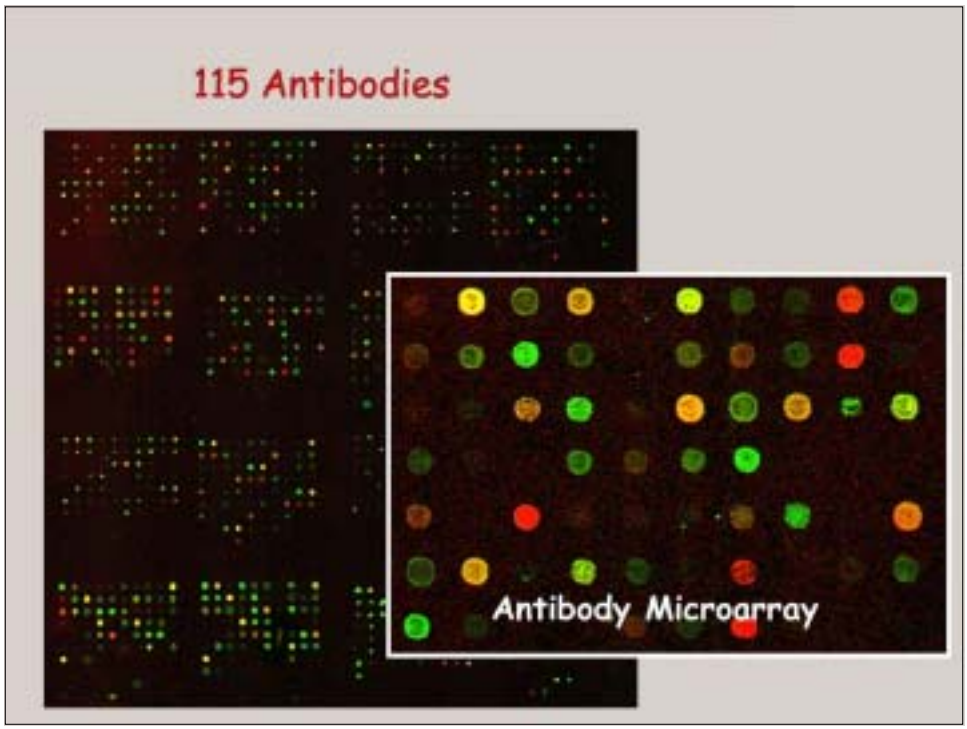




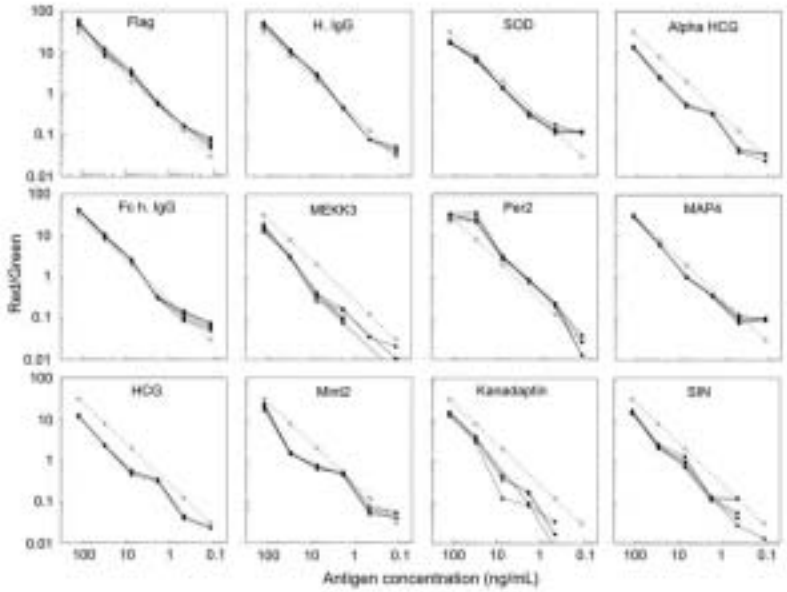
Hybridization of labeled Genomic DNA results in the measurement of DNA copy number and can be used to map amplifications and deletions with sub gene resolution



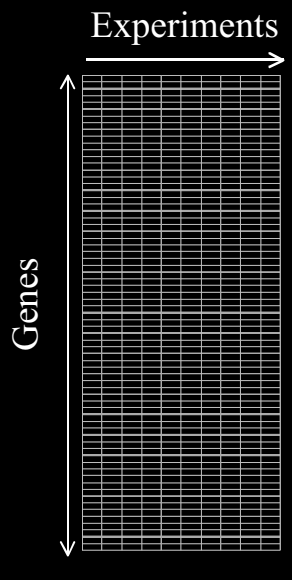

a
b
c
moving average
number of X chromosomes
number of X chromosomes
fluorescence ratio
fluorescence ratio
autosomal cDNAs
X-chromosomal cDNAs


115 Antibodies
Antibody Microarray

Antibody Microarray Performance



The Principle Challenge in Experimental Genomics is Making Biological Sense of the Data

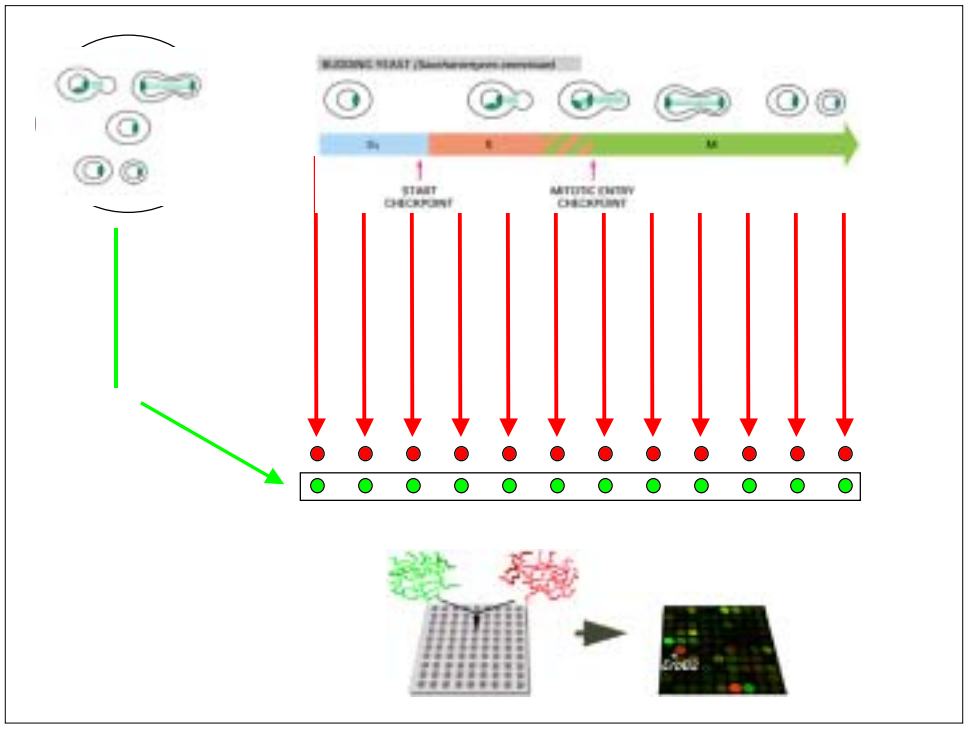


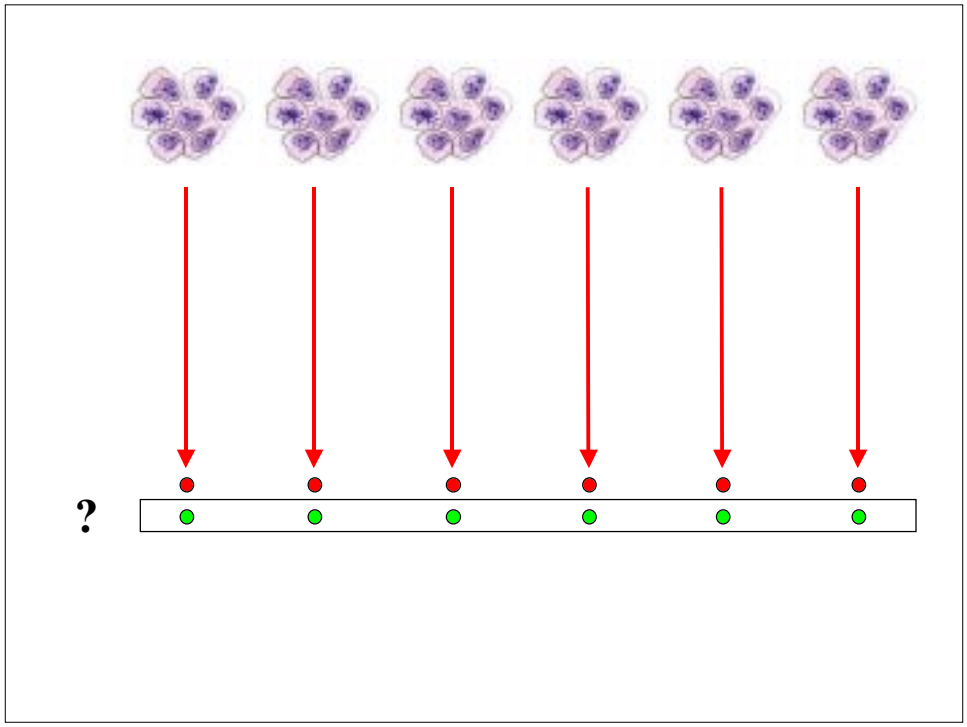
Saccharomyces cerevisiae
whole genome ORF arrays

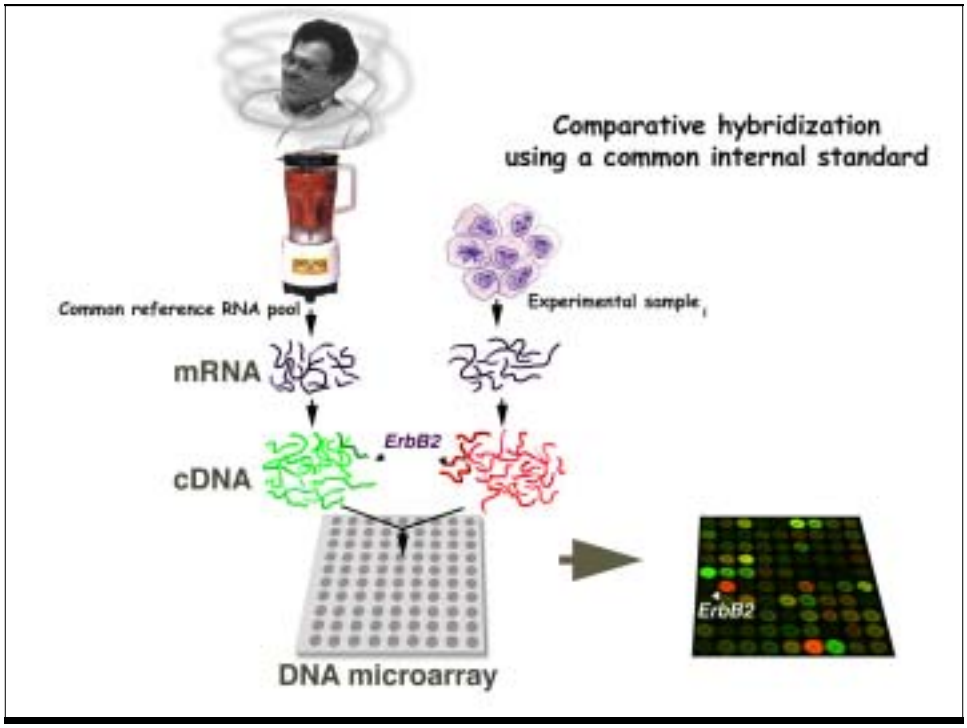
~6200 genes
~6000 conditions
> 36 million observations

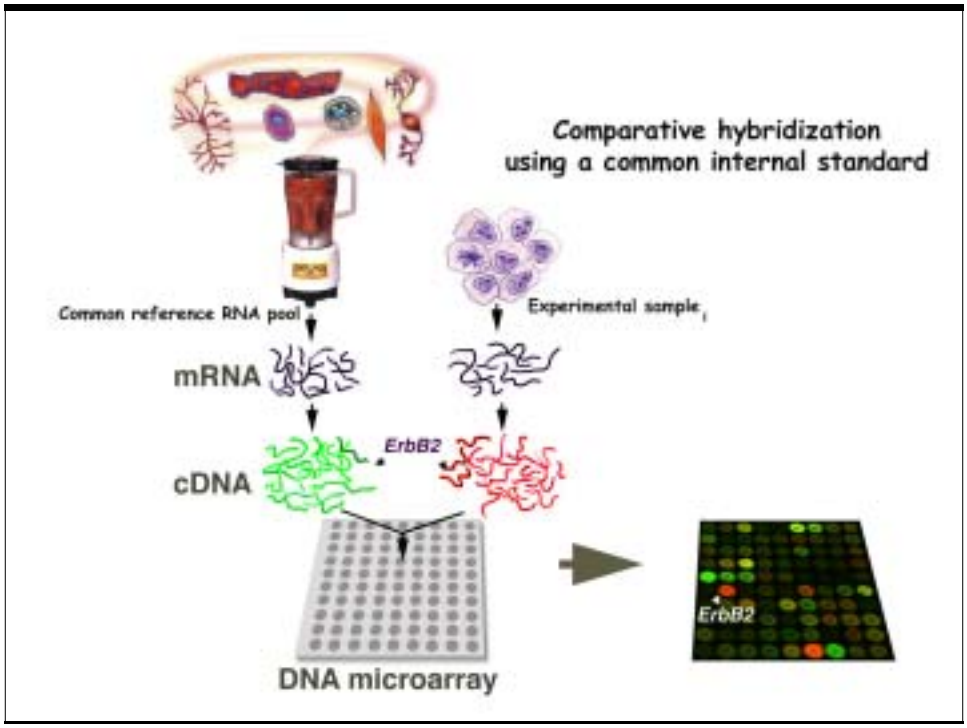
Human
cDNA arrays

10,000 - 25,000 genes
~4000 conditions
>75 million observations


START CHECKPOINT
MITOTIC ENTRY CHECKPOINT


?


Comparative hybridization
using a common internal standard
Common reference RNA pool
Experimental sample
mRNA
cDNA
ErbB2
DNA microarray
ErbB2


Comparative hybridization
using a common internal standard
Common reference RNA pool
Experimental sample
mRNA
cDNA
ErbB2
DNA microarray
ErbB2

Why Do We Measure Gene Expression?

Why Do We Measure Gene Expression?

Gene expression – which is now easy to measure on a genomic scale – provides a window on aspects of biology in which we are interested that are difficult to measure directly.

Why Do We Measure Gene Expression?

Natural selection ensures that genes code for proteins with the proper molecular properties (enzymatic activity, binding, …).

Why Do We Measure Gene Expression?

Natural also acts on gene expression patterns.

It is also important that proteins be made when, where and in the proper amounts required, that they are not made when their presence would be deleterious, and that cellular energy is not wasted synthesizing unnecessary genes.

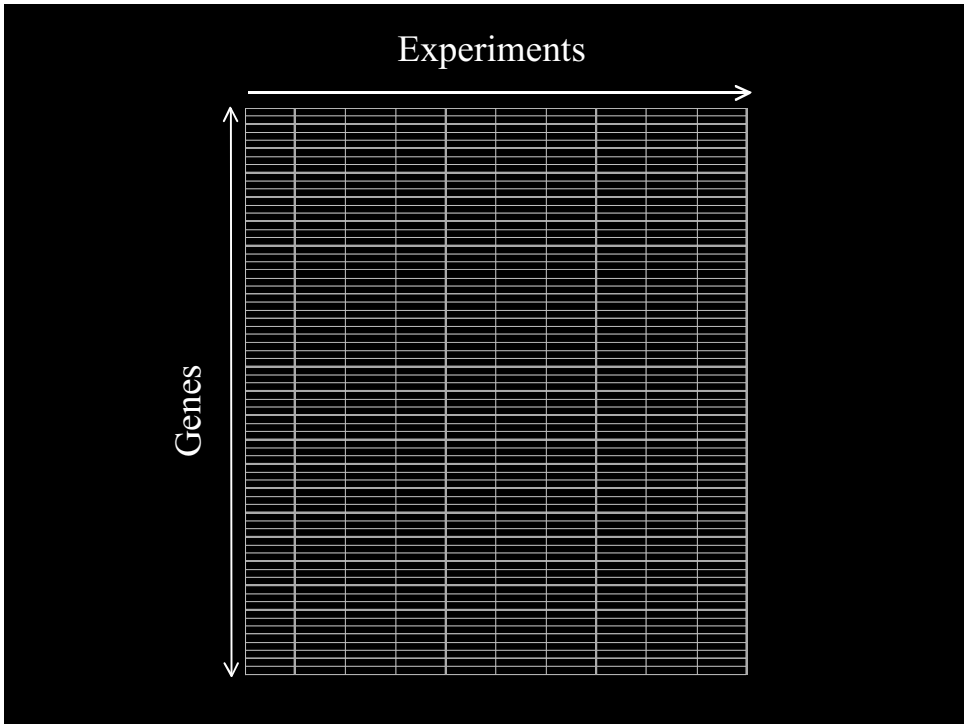
Why Do We Measure Gene Expression?

Since the particular set of condition in which a protein is required, and the particular amount of it that is required in these conditions, reflects to molecular function and cellular role of the protein, there is a logical connection between protein function and gene expression patterns.

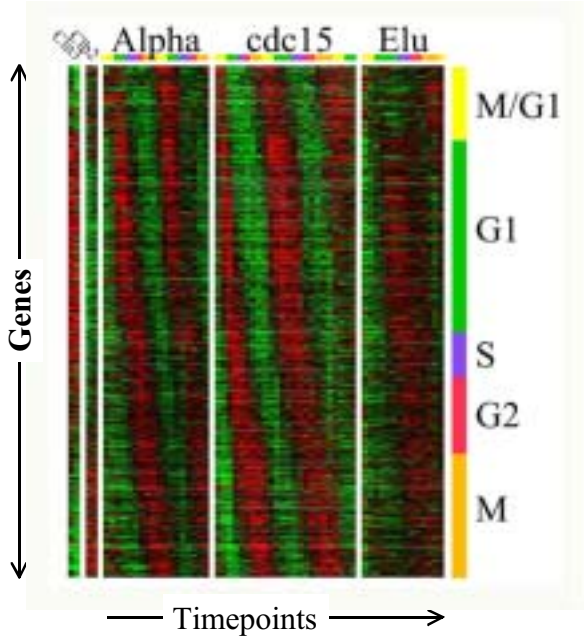
The Relationship Between Gene Function and Gene Expression Patterns

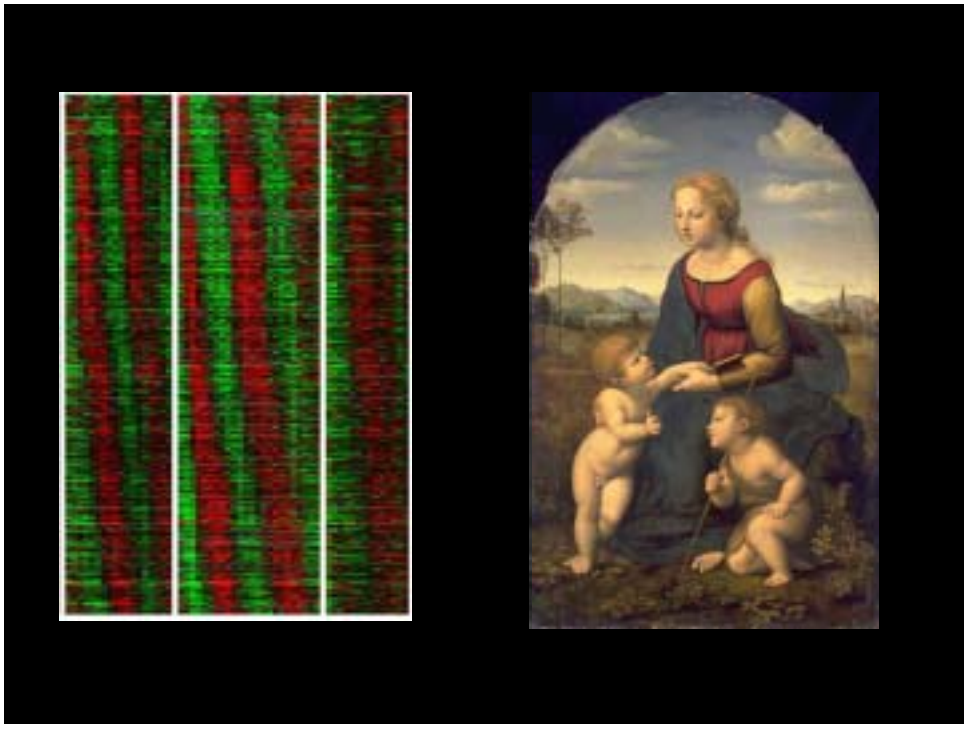
A corollary of this logical relationship between gene expression and gene function is that genes with similar function should have similar patterns of gene expression.

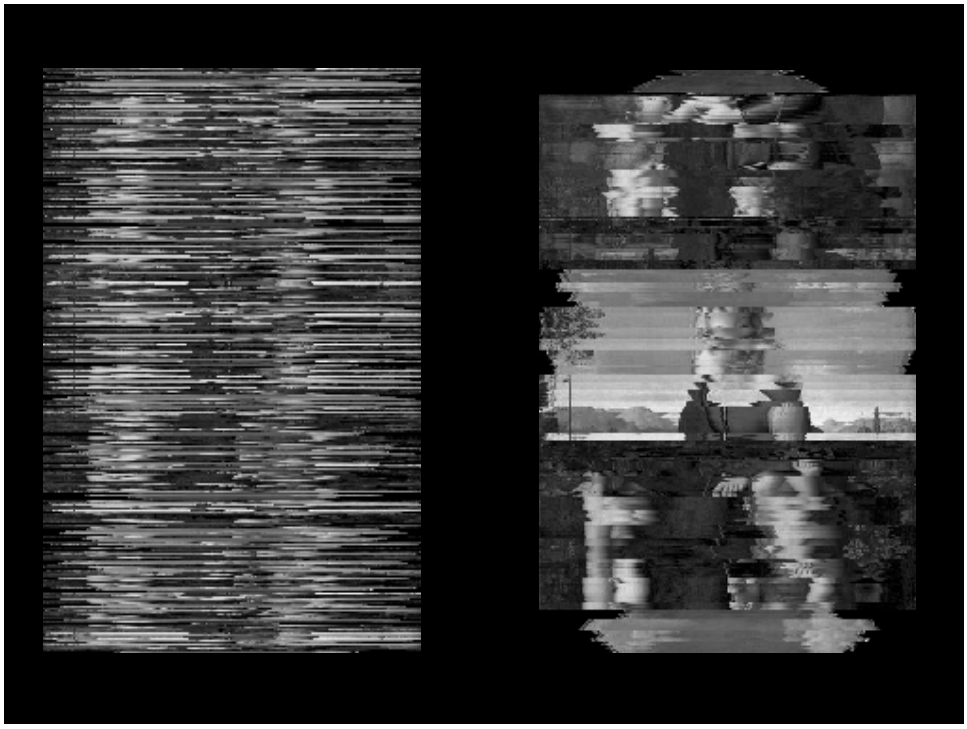
To the extent that this is true, this property can be used to impart a logical and biologically meaningful order to complex gene expression data

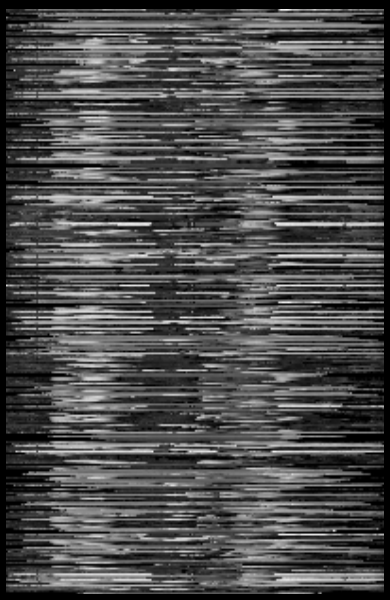


A simple example: The mitotic cell-cycle in yeast







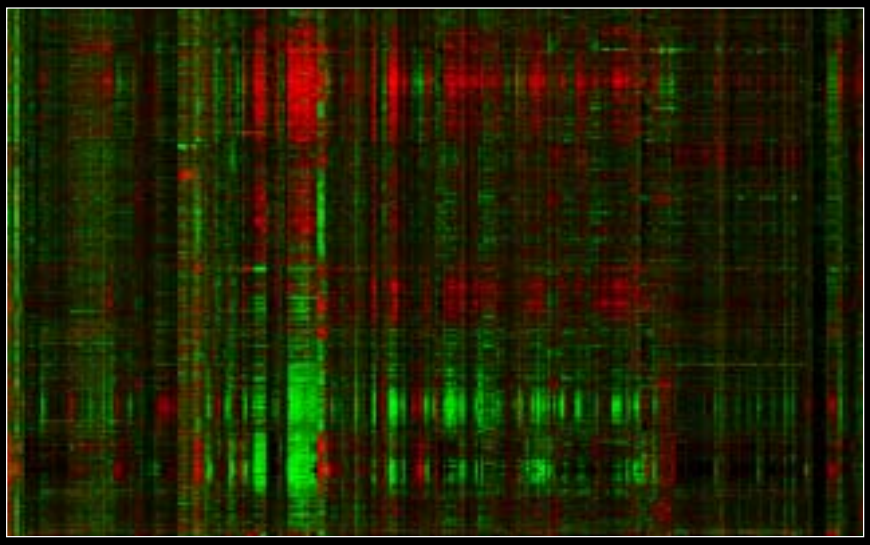


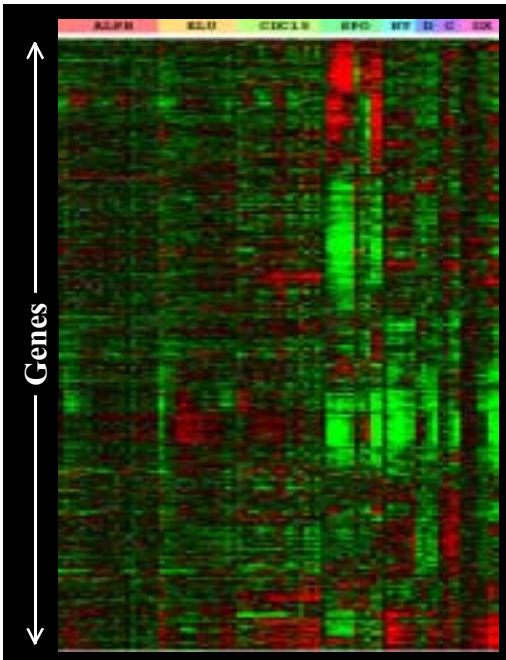
Clustering and Related Methods Applied to Gene Expression Data

Hierarchical Clustering (many flavors)
Self-organizing maps
K-means
DIANA
Support vector machines
Gene shaving
Principal Component Analysis
Multi-dimensional Scaling/Sammon Mapping
…

and there exists a large armory of other methods waiting simply for someone to type (or copy) some code

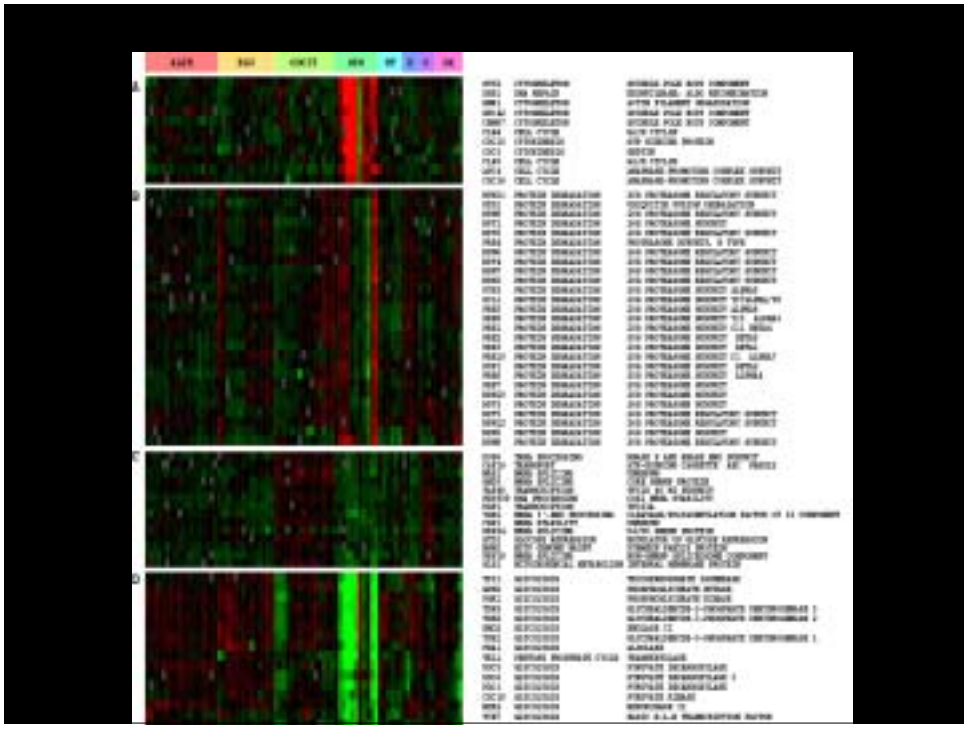
Gene Expression Program of *Saccharomyces cerevisiae*

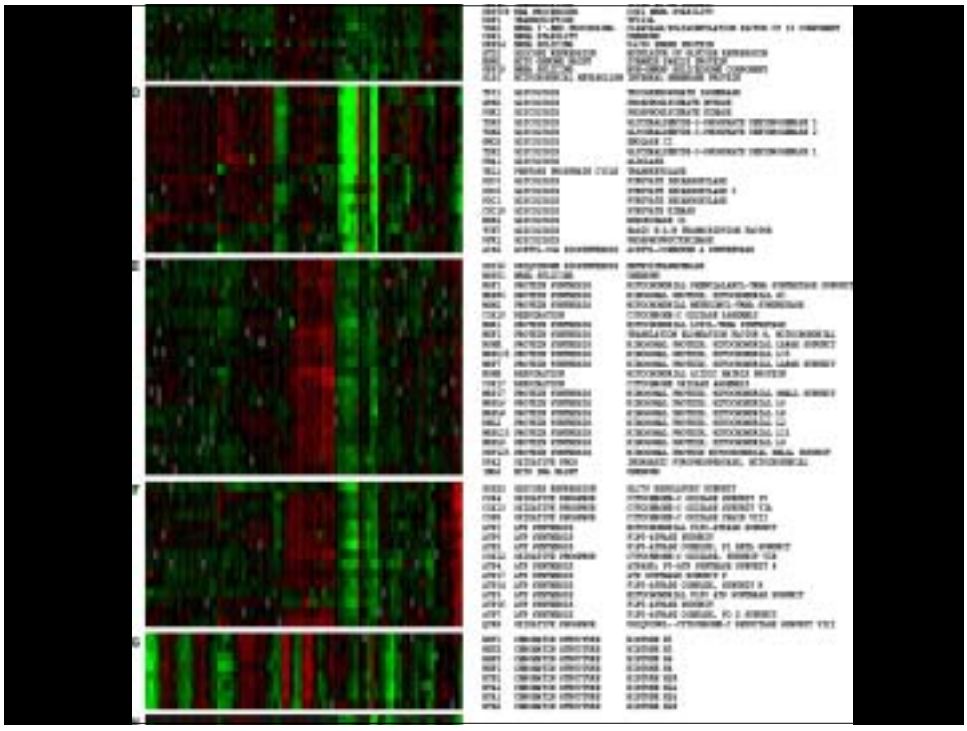


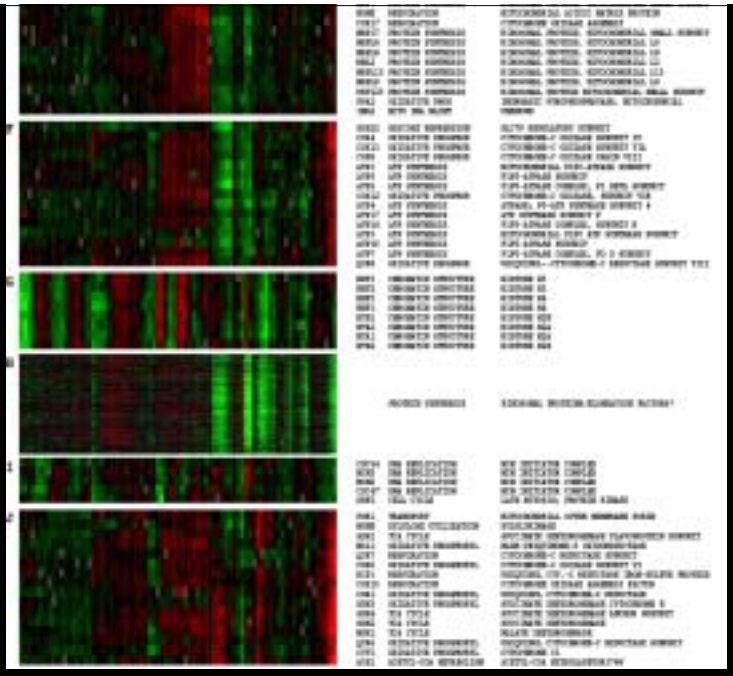


S. Cerevisiae

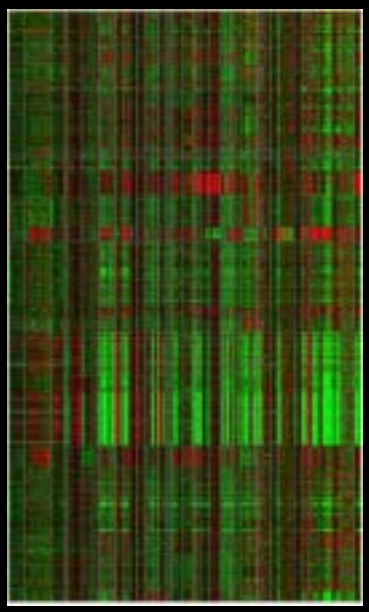
Gene Expression During Important Biological Processes



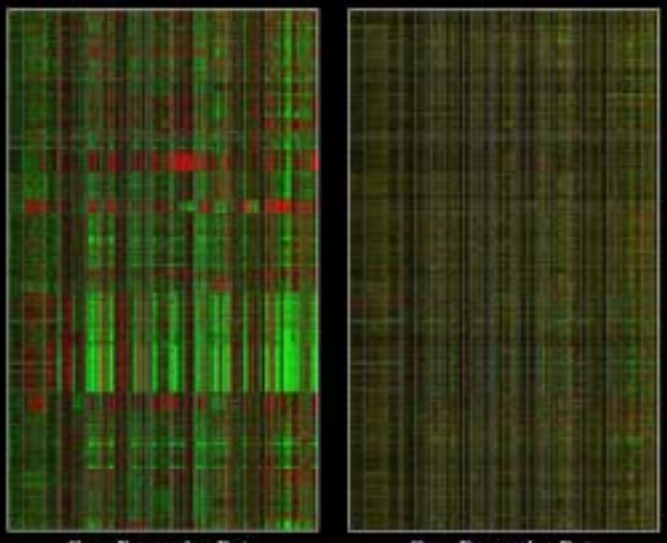




S. cerevisiae multi-subunit complexes



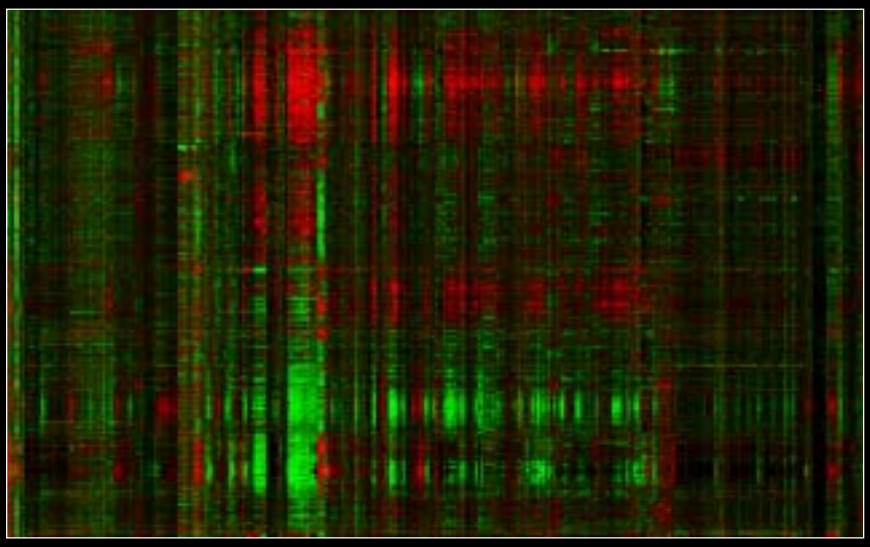
S. cerevisiae multi-subunit complexes



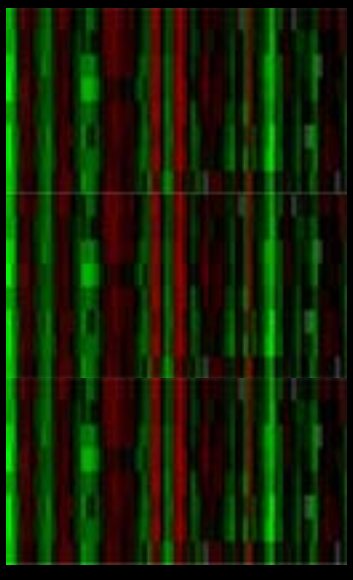
Gene Expression Data

Gene Expression Data
orthogonalized within each complex

Relating Gene Expression and Genome Sequences



Identification of cis-regulatory sites from gene expression data



Possible examples of motif 1 in the training set

Sequence name	strand	Start	Score	Site
YLR118C	w53	736	10.74	AAATAAAAAT GTTAGCAAAA GAGAACGAAA
YDL126C	w53	660	17.33	ATGTGAATTC GGTGGCAAAA AAGGACATAT
YER021W	c53	307	10.44	TTTCTTTGAT GGCGGCAAAT ACTAATCTTA
YER021W	c53	705	17.07	CTCTTTCTTG GGTGGCAAAT GTTTTTCGTC
YPR108W	c53	646	15.17	ATCCTCAATG GGTGGCAAAC TTAACTTATT
YDR427W	w53	560	9.77	TCACGCATCG TGTTGCAAAT AGCATATACA
YDR427W	w53	692	17.33	GCTTTCATCC GGTGGCAAAA GTAAGAACAA
YOR157C	w53	643	14.30	AACCCGGCCG AGTGGCAAAA TACCAAAAAG
YDL097C	w53	688	17.07	TGGTTCCAAG GGTGGCAAAT GTTGGTAGAA
YOR117W	w53	637	17.33	GCCAATATGT GGTGGCAAAA ATGAATTAAT
YML092C	w53	681	17.33	AGCTTATTAA GGTGGCAAAA TTGTCCCCGG
YER012W	w53	659	17.33	AAATCTTTAC GGTGGCAAAA AATAAAGAAA
YLR387C	w53	709	15.03	GTCACAAGAC GGTGGCAACA CAGTTGTACA
YDR394W	w53	625	17.33	CTGAAAGTAA GGTGGCAAAA TAGTAAATCT
YPR103W	w53	646	17.33	ACAGCAAAAG GGTGGCAAAA CGAAGAATAG
YJL001W	w53	683	17.33	AACGGAATCC GGTGGCAAAA AAGGGAAAAG
YFR050C	c53	689	17.07	ACGTATTTAA GGTGGCAAAT TCTAAACCAA
YFR052W	c53	703	17.07	AATTTCAATG GGTGGCAAAT AATGTGTAGA
YGL011C	w53	446	10.40	TTAGGATTGT GGTGGTAAAA CTTGTAAAAT
YGL011C	c53	692	17.07	CTATCCTTAC GGTGGCAAAT TGGTGTGTGC
YKL145W	c53	645	17.33	TTAGTTTTCG GGTGGCAAAA TTCGGTAATC
YGL048C	w53	717	17.33	ATGCGAGCGA GGTGGCAAAA GTTGTAATAT
YGR135W	w53	642	17.33	TTACCCTGCA GGTGGCAAAA AATTTATAGA
YGR253C	c53	480	10.28	CTTCAGCAGT AGTTGCAAAA GTGTCCTTTT
YGR253C	w53	697	17.33	AACCAAGACC GGTGGCAAAA GAGTAACCTA
YBL041W	w53	714	17.33	AAACTCAAAG GGTGGCAAAA ACAAGCGGTA
YFR004W	w53	677	17.07	GGACCTCCAA GGTGGCAAAT TCACAAGAAA
YJL031C	c53	162	11.38	GTGGCTTCTT GGTGGAAAAT GCAATCAATT
YJL031C	c53	675	13.79	TGCTGTTTAA TGTGGCAAAT ATAAAAAATT
YMR004W	w53	169	12.36	AAAATTATCA GGTCGCAAAT CAGATTTGCT
YMR004W	c53	182	10.21	GATAAGTCTT TGTAGCAAAT CTGATTTGCG
YBR170C	w53	725	14.03	CAGAACGAAG AGTGGCAAAT AATAGGACAG
YOR059C	w53	727	14.77	ATAAAGAAAG GGTGGCAACT CTTTACTGCA
YKR014C	w53	168	10.78	AATGTGGATT TTTGGCAAAT CAATGTTTCT
YJL053W	c53	465	10.08	GACCCAGTTC TGTGGCCAAT CTTTTTCTTC
YPR107C	w53	557	15.17	ATCCTCAATG GGTGGCAAAC TTAACTTATT
YGL047W	c53	113	11.64	CGTTGATGAT GGTGGAAAAA GTTCCTGATT
YGL047W	c53	236	10.04	ACATCGTGGA TGTTGCAAAA GATATAAACG
YGL047W	c53	610	17.33	ATGCGAGCGA GGTGGCAAAA GTTGTAATAT

RPN4 Subunit of the regulatory particle of the proteasome

Gene Name/Synonyms RPN4; SON1; UFD5; GVM1; D2840; YDL020C

At-a-Glance

Cellular Role	Protein degradation [details]
Biochemical Function	Proteasome subunit [details]
Localization	Nuclear;19S regulatory particle of the proteasome [details]
Mutant Phenotype	Null: viable [details]

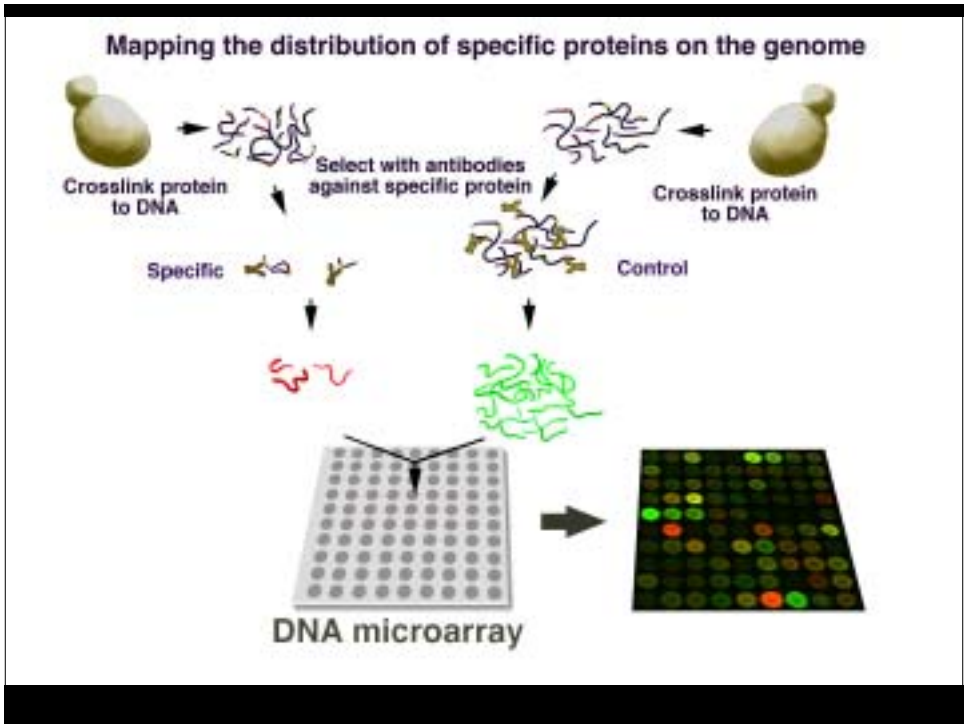
mammalian homolog cannot be found in purified proteasomes

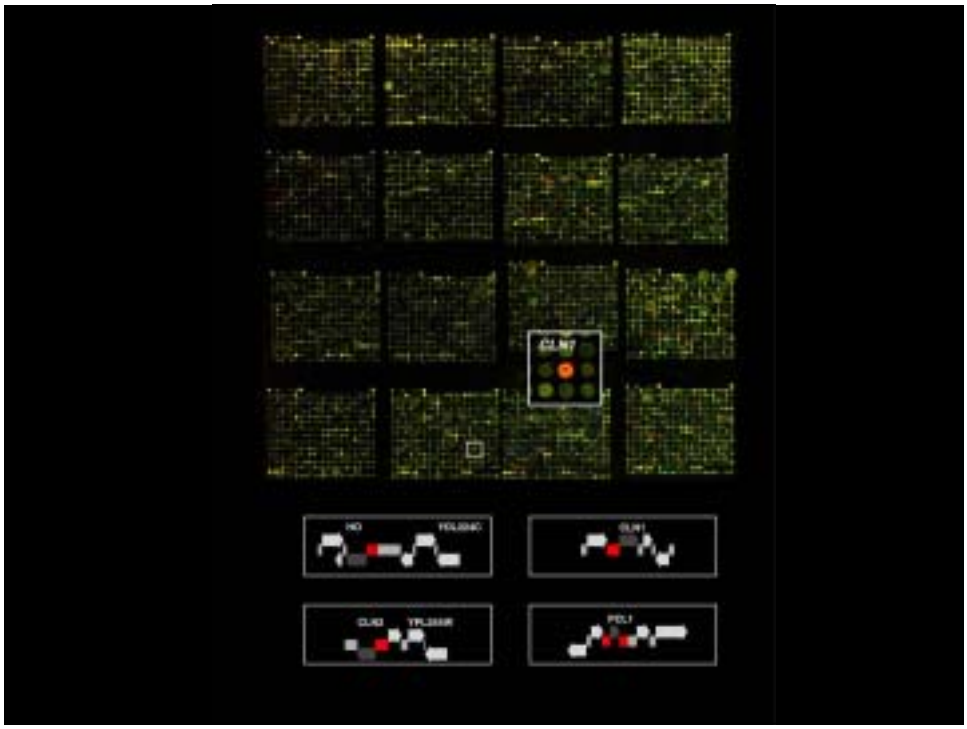
has two potential nuclear localization (NLS) sequences

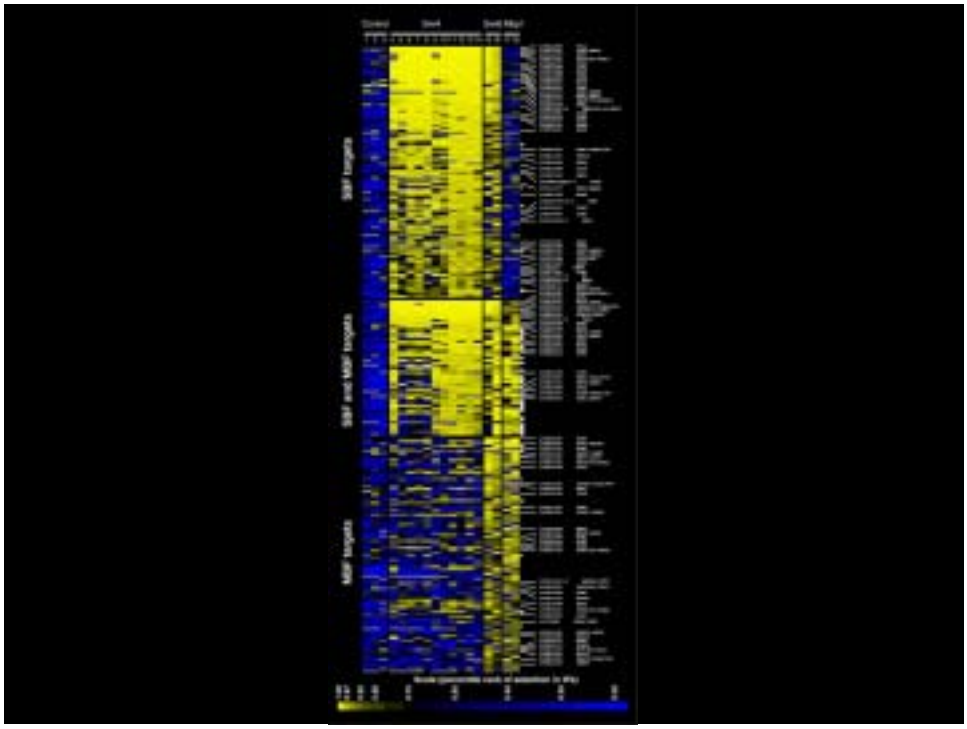
not detectable in the purified proteasome preparation by direct sequencing or by detection with antibodies

Expression Data and Genome Sequences

- Many nice methods have been developed to identify sequence fragments shared by co-expressed genes with much lower than random probability, and many such sequence fragments can be found.
- Far fewer sites have been found than must exist.
- Even the best sites are not found in all genes in cluster and are found in many other genes that have very different expression patterns.
- It is clear that we need a better understanding of the rules that govern in vivo binding and the way in which in vivo binding of txn factors is read out as txn rates.


Mapping the distribution of specific proteins on the genome
Crosslink protein to DNA
Select with antibodies against specific protein
Crosslink protein to DNA
Specific
Control
DNA microarray





Ongoing Projects

- Public database of gene expression in yeast over 100,000 growth condition
- Description and characterization of yeast genes based on expression patterns
- Understanding logic and mechanisms of transcriptional regulation by combining gene expression data, sequence analysis, evolutionary comparison and ChIP (yeast, Drosophila, mouse)
- Genetic mapping of heritable complex gene expression phenotypes in wild yeast

Genomics and Information Access

- Genomics, and science in general, will depend increasingly on ready and unfettered access to data and information
- Public databases of gene expression and other genomic data are vital.
- For these to be optimally useful, the intellectual capital produced by scientists stored in the archived and future scientific literature must also be completely in the public domain

Towards a Public Library of Science

- Ensure that the only public record of the scientific process belongs to the public by pledging not to publish in, edit for or review articles from journals that do not agree to release all of the research reports they publish into the public domain within 6 months of its publication
- **www.publiclibraryofscience.org**

Acknowledgements

David Botstein
Pat Brown

Joe DeRisi
Ash Alizadeh
Charles M. Perou
Doug Ross
Max Diehn
Audrey Gasch
Paul Spellman
Gavin Sherlock
Christian Rees
Sandrine Dudoit

Vishy Iyer (Austin)
Mike Snyder (Yale)
Christine Horak (Yale)

Robert Tibshirani
Trevor Hastie
Izidore S. Lossos
James O. Armitage (UNMC)
Ronald Levy
Lou Staudt (NCI)
Matt van de Rijn
Stefanie S. Jeffrey
Therese Sørlie (NRH)
Anne-Lise Børresen-Dale (NRH)
Gyorgi Dendroica (HHF)

Funding and Support

Alfred P. Sloan Foundation
National Cancer Institute
Howard Hughes Medical Institutes
Naval Space Warfare Administration
Lawrence Berkeley National Lab

Axon Instruments
Research Genetics

www.microarrays.org

Sofware

rana.lbl.gov

Eisen Lab Website

llmpp.nih.gov

genome-www.stanford.edu